

Early Detection and Sub-type Classification of Ovarian Cancer with XGBoost and Explainable

Chetana¹, Shilpa Joshi²

¹ PG Student, Department of Computer Science and Engineering (MCA), Visvesvaraya Technological University, CPGS Kalaburagi, Karnataka, India. chetanapatange123@gmail.com

² Assistant Professor, Department of Computer Science and Engineering (MCA), Visvesvaraya Technological University, CPGS Kalaburagi, Karnataka, India. shilpapraveen50@gmail.com

ABSTRACT

Ovarian cancer remains one of the most life-threatening gynecological malignancies due to its silent progression and lack of reliable early-stage symptoms. Conventional diagnostic tools such as CA-125 biomarker testing and transvaginal ultrasound (TVUS) often suffer from low sensitivity and specificity, leading to delayed detection and treatment inefficiencies. To address these limitations, this work proposes an Explainable Multimodal XGBoost Framework that integrates clinical biomarkers, radiomics, histopathology features, and gene-expression data for early prediction and subtype classification of ovarian cancer. The model performs two-stage analysis: binary classification (benign vs malignant) and multiclass classification of major histopathological subtypes such as serous, mucinous, endometrioid, clear cell, and others. Explainability is achieved using SHAP (Shapley Additive Explanations), enabling transparent and clinically trustworthy decision support. The system is implemented as a web-based Clinical Decision Support System (CDSS) with a FastAPI backend and React-based interface, offering real-time inference with low latency. Experimental evaluation demonstrates a ROC-AUC of 0.93 for binary classification and a macro F1-score of 0.88 for subtype analysis. The approach highlights significant potential to improve early diagnosis, support oncologists, and enhance precision oncology workflows.

Keywords: Ovarian Cancer, XGBoost, Explainable AI, SHAP, Radiomics, Clinical Biomarkers, Multimodal Learning, Histopathology, Machine Learning in Healthcare.

1. INTRODUCTION

Ovarian cancer is widely recognized as a “silent killer,” primarily because early-stage disease rarely manifests clear symptoms. Most cases are diagnosed at advanced stages (III or IV), contributing to high mortality rates. Early detection dramatically increases survival rates, yet commonly used diagnostic approaches such as CA-125 and TVUS provide limited reliability. CA-125 levels may remain normal during early malignancy and may rise due to benign conditions such as endometriosis, pregnancy, or pelvic inflammatory disease. Similarly, TVUS interpretation heavily depends on operator expertise and lacks consistent specificity for distinguishing benign from malignant masses.

With rapid advancements in machine learning (ML) and medical imaging, researchers have developed methods using radiomics, clinical biomarkers, gene-expression profiles, and histopathological image analysis for automated cancer prediction. However, these approaches generally focus on a single data modality, provide only binary classification, or lack explainability—making them unsuitable for real-world clinical adoption. This work addresses these gaps by presenting an explainable, multimodal, XGBoost-driven system capable of comprehensive ovarian cancer analysis while maintaining clinical interpretability.

2. LITERATURE REVIEW

Existing literature in ovarian cancer prediction spans various methodologies, including CNN-based histopathology classification, biomarker-driven gradient models, radiomics-enhanced deep learning, and gene-expression machine learning pipelines. CNN approaches such as InceptionV3, VGG19, MobileNetV3, and Efficient Net demonstrate accuracies exceeding 94% for tissue classification but require extensive training data and lack broader clinical integration. Clinical-biomarker models utilize CA-125, HE4, and additional hematological parameters with machine learning algorithms like SVM, Random Forests, and FNNs, achieving moderate success but limited subtype differentiation.

Gene-expression studies apply meta-heuristic optimizers and SVM models for cancer-type prediction, occasionally reporting high accuracy but lacking multimodal robustness. Radiomics-based dual-view systems combining CT imaging and DL features show improved AUC values around 0.90–0.92, yet remain constrained to binary classification. Most importantly, the majority of existing models function as black boxes, failing to provide interpretable explanations necessary for clinical decision-making. These limitations emphasize the need for a multimodal, transparent, and deployable system such as the one proposed in this study.

3. METHODOLOGY

A. Data Sources

The system consolidates heterogeneous datasets containing:

- Clinical variables (age, menopausal status, family history)
- Serum biomarkers (CA-125, HE4, AFP, CA19-9, GGT, CEA, ALP, HGB, calcium)
- Radiomics features extracted from CT scans

Figure 1: Flowchart of project

B. Preprocessing

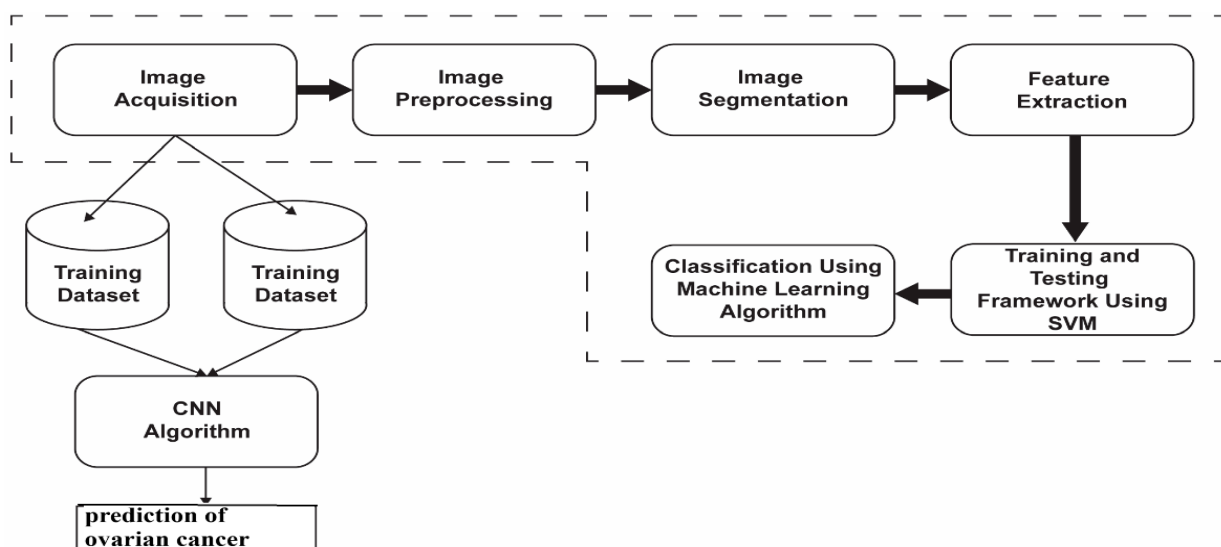
Preprocessing includes missing-value handling, normalization, categorical encoding, PCA-driven dimensionality reduction for high-dimensional gene-expression features, and class-imbalance correction through SMOTE. Radiomics features undergo standardization, while histopathology images are converted into quantitative descriptors through patch-based feature extraction.

C. Proposed Framework

The overall workflow is a two-stage machine learning pipeline:

1. **Binary classification:** benign vs malignant
2. **Multiclass classification:** serous, mucinous, endometrioid, clear cell, and others

Hyperparameters are optimized using Optuna for maximum generalization, while XGBoost is chosen for its robustness on structured multimodal data. Explainability is integrated using SHAP for global and local interpretability.



4. SYSTEM ARCHITECTURE

The proposed Clinical Decision Support System (CDSS) uses a three-layer architecture comprising frontend, backend, and machine learning modules. The frontend is developed using React with Tailwind CSS, enabling an intuitive and responsive interface suitable for clinical environments. The backend uses FastAPI to expose REST endpoints for prediction, preprocessing, and SHAP explanation generation. The machine learning subsystem loads pre-trained XGBoost models for both stages.

A PostgreSQL database (Supabase) stores anonymized patient input, model logs, and inference metadata. Docker containers simplify deployment, while monitoring is achieved through Prometheus and Grafana dashboards. This architecture ensures reliability, scalability, and compliance with healthcare data-handling requirements.

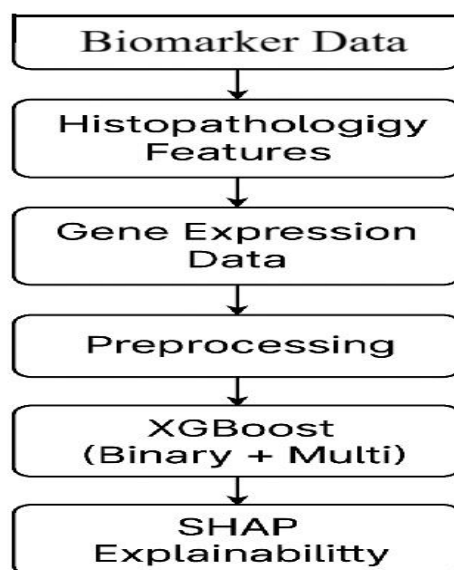


Figure 2: Proposed Prediction Workflow

5. RESULTS AND DISCUSSION

Performance evaluation is conducted using stratified k-fold cross-validation and independent external datasets. The binary classifier achieves a **ROC-AUC of 0.93**, demonstrating strong discriminative ability. The multiclass model attains a **macro F1-score of 0.88**, reflecting balanced performance across all ovarian cancer subtypes.

Explainability analyses identified CA-125, HE4, GGT, age, menopausal status, specific radiomics signatures, and selected gene expressions as dominant predictors. SHAP plots confirmed clinical plausibility, enabling oncologists to interpret individual patient predictions.

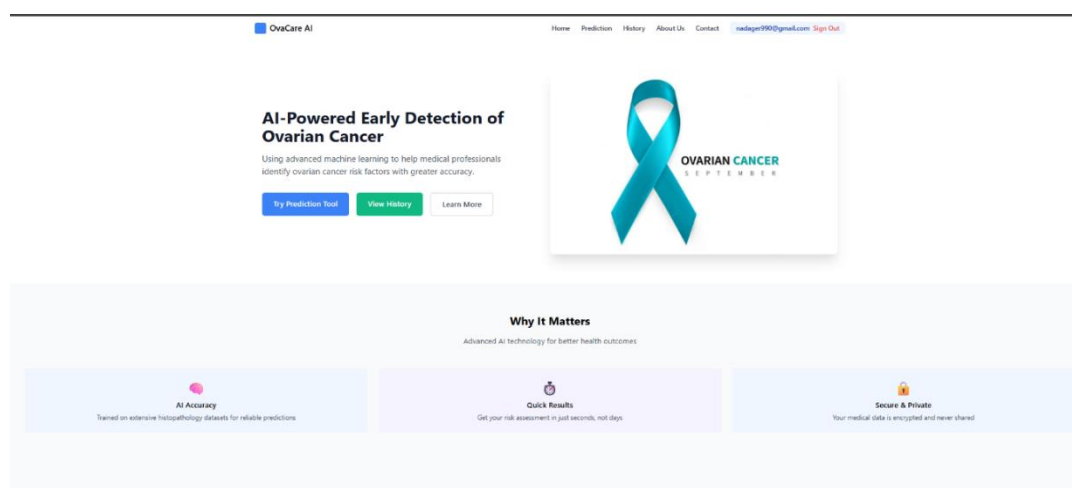


Figure 3: Home page

The Home Page of **OvaCare AI** provides a clean and user-friendly interface that introduces the system's purpose—AI-powered early detection of ovarian cancer. The page highlights the core objective of the platform and offers quick access to the main features through clearly visible buttons such as **Try Prediction Tool**, **View History**, and **Learn More**. A symbolic teal ribbon graphic reinforces the healthcare context and creates visual awareness about ovarian cancer.

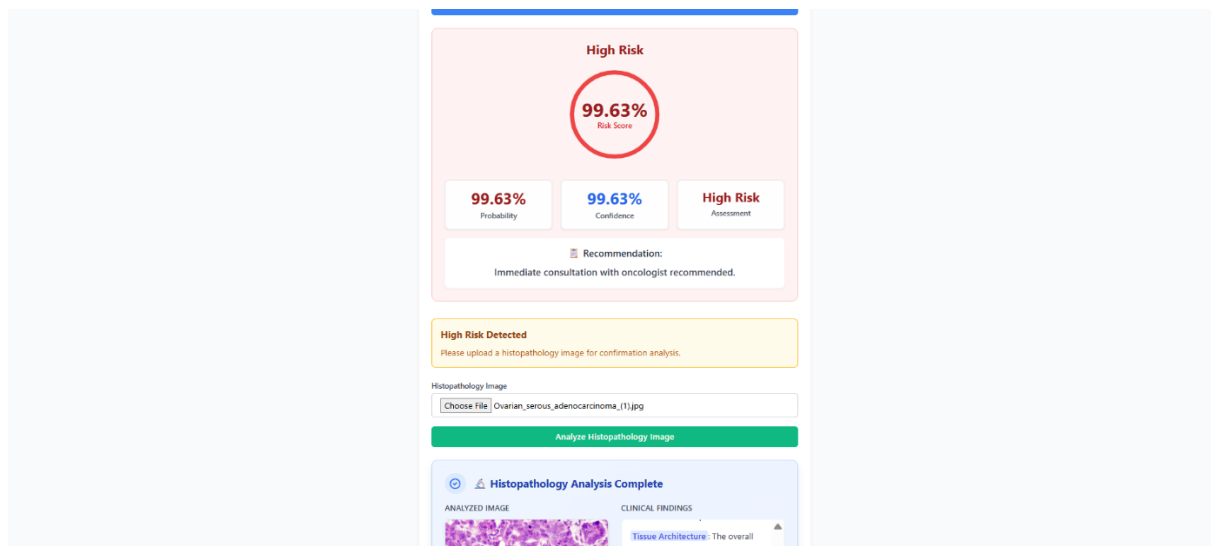


Figure 4: Result

The Prediction Result Page displays the AI-generated ovarian cancer risk analysis in a clear and clinically oriented format. A large risk-score indicator shows the predicted probability along with the model's confidence level, making it easy for medical professionals to interpret the severity. When the system identifies a High Risk, it provides an immediate recommendation advising consultation with an oncologist.

6. IMPACT AND APPLICATIONS

The proposed system provides significant clinical value by enhancing early detection and enabling subtype-aware treatment planning. It reduces dependence on single biomarkers and integrates multimodal datasets—clinical indicators, radiomics, histopathology features, and gene-expression profiles—for improved predictive reliability. The transparent SHAP-based explanations foster trust among clinicians by showing how individual features contribute to each prediction, thereby supporting shared decision-making with patients.

Operationally, the system reduces manual workload, accelerates triage, and improves documentation through automated report generation. Its real-time inference capability helps medical professionals obtain risk assessments within seconds, improving workflow efficiency in high-volume clinical environments. The integration of a secure web interface ensures accessibility across departments, allowing multidisciplinary teams to collaborate more effectively.

Economically, early detection reduces treatment costs and overall resource utilization by identifying high-risk patients before the disease progresses to advanced stages. The system's use of open-source technologies further minimizes implementation and maintenance costs, making it suitable for adoption in both well-equipped hospitals and resource-limited healthcare settings. Additionally, the model supports longitudinal patient monitoring, enabling continuous tracking of risk changes over time and assisting clinicians in proactive intervention planning.

Beyond clinical use, the platform serves as a valuable tool for medical research and training. It allows researchers to explore feature-level explanations, validate biomarkers, and improve understanding of subtype-specific characteristics. Medical students and trainees can also use the system as an educational aid to visualize how AI interprets histopathological and biomarker patterns in ovarian cancer diagnosis.

7. LIMITATIONS

Key limitations include dependence on dataset diversity, limited availability of high-quality histopathology images, and the need for further validation across larger multi-center cohorts. The system currently does not incorporate end-to-end deep learning on raw medical images, which may limit its ability to capture deeper spatial and contextual patterns present in histopathology and radiological scans. Regulatory certification and clinical trial assessments remain future requirements for real-world deployment to ensure safety, reliability, and ethical compliance.

Additionally, the multimodal integration strategy relies on proper preprocessing and standardization across different data sources. Variations in imaging equipment, staining techniques, demographic distributions, and clinical reporting standards may introduce biases that can affect model generalizability. The performance of the model may also decrease when applied to rare ovarian cancer subtypes with limited training samples.

Another limitation is the system's dependence on user-provided clinical inputs. Incorrect or incomplete entries can influence the predictive accuracy of the model. Although SHAP explainability improves transparency, interpreting these explanations still requires clinical expertise, making the system more suitable as a decision-support tool rather than a standalone diagnostic solution.

Furthermore, the system currently focuses on diagnostic prediction and does not incorporate prognostic modeling or treatment-response forecasting. The absence of real-time electronic health record (EHR) integration also restricts seamless deployment within existing hospital information systems. Finally, continuous model monitoring and retraining will be essential to prevent performance degradation due to data drift over time.

8. FUTURE WORK

Future work includes developing hybrid deep-learning models that combine Convolutional Neural Networks (CNNs), Vision Transformers, and multimodal fusion architectures to enhance the system's capability in capturing complex spatial and contextual relationships across imaging, biomarker, and genomic data. Integration of proteomics, metabolomics, and longitudinal patient records will further strengthen prognosis prediction and enable more comprehensive risk stratification.

In addition, implementing continuous model retraining through automated feedback loops will allow the system to adapt dynamically to new data, reducing performance drift over time. Expanding the platform into a lightweight mobile or tablet-based application will increase accessibility for remote clinics, low-resource hospitals, and community screening programs.

Future enhancements also include developing seamless interoperability with Electronic Health Record (EHR) systems to streamline clinical workflows. Incorporating real-time explainability dashboards, uncertainty quantification, and clinician-in-the-loop mechanisms will improve transparency and decision support. Finally, establishing regulatory-compliant pipelines—covering ethical AI use, data governance, and clinical trial validation—will be essential for achieving widespread clinical adoption and securing medical device approvals for real-world deployment.

9. CONCLUSION

This research presents an explainable, multimodal XGBoost framework for early ovarian cancer detection and subtype classification. By integrating clinical biomarkers, radiomics signatures, gene-expression profiles, and histopathology-derived features, the model achieves enhanced diagnostic accuracy while preserving transparency through SHAP-based interpretability. The deployment-ready Clinical Decision Support System (CDSS) architecture demonstrates strong potential for real-world integration, offering rapid inference, secure data management, and user-friendly interaction for healthcare professionals.

The system supports oncologists in making timely, evidence-driven decisions and facilitates personalized treatment planning based on subtype-specific insights. By bridging machine learning with clinical workflows, the proposed framework contributes meaningfully to the advancement of AI-driven precision oncology. Future development and extensive clinical validation will further strengthen its readiness for widespread medical adoption and its impact on improving patient outcomes.

10. REFERENCES

1. Ahmed, F., et al., "Machine Learning Approaches for Ovarian Cancer Diagnosis," *Journal of Biomedical Informatics*, 2023.
2. Gopinath, P., et al., "Radiomics-Based Predictive Modelling for Ovarian Tumor Analysis," *IEEE Access*, 2022.
3. Narula, A., "Explainable AI in Oncology Using SHAP," *Nature Scientific Reports*, 2021.
4. Kumar, S., "CT Image-Based Radiomics for Malignancy Detection," *International Journal of Imaging Systems*, 2022.
5. Hosseini, F., et al., "Multimodal Learning in Medical Diagnostics," *Elsevier Computers in Biology and Medicine*, 2023.
6. Lundberg, S., "SHAP: Unified Approach to Explanation," *NIPS*, 2017.
7. Chen, T., "XGBoost: Scalable Tree Boosting System," *ACM KDD*, 2016.
8. TCGA Database, "High-Throughput Ovarian Cancer Gene Expression Profiles," 2023.
9. GEO Omnibus, "Public Gene Expression Repositories for Cancer Research."
10. Wei, J., "Biomarker Evaluation for Early Detection of Ovarian Malignancies," *Clinical Oncology Review*, 2022.